

Collaborative Classifier Agents: Studying the Impact of Learning in Distributed Knowledge Discovery

Weimao Ke, Javed Mostafa, and Yueyu Fu
Laboratory of Applied Informatics Research
{wke, jm, yufu}@indiana.edu

ABSTRACT

Classification, a knowledge organization mechanism, is a way we humans understand the world by aggregating like-entities (Lewis, 1992; Yang, 2002). According to Taulbee (1965), "the ability to classify is an essential part of life." Text Classification, or Categorization, has been a research area in Machine Learning (ML) and Information Retrieval (IR) and is relevant to information extraction and knowledge discovery (Knight, 1999; Sebastiani, 2002). It is a fundamental function of IR and can be applied to various processes such as indexing and filtering (Taulbee, 1965; Mostafa & Lam, 2000). Traditional classification approaches assume that global knowledge is available at a centralized place. However, this assumption is rarely true in the real world. Evidence is emerging that, given the distributed nature of knowledge/information, collaboration is the main force driving knowledge management/discovery and making possible the emergence of a global brain (Wimmer & Traummuller, 2000; Red'ko, 2002; Börner, Dall'Asta, Ke, & Vespignani, 2005). The World Wide Web is a good example of knowledge/information distribution, where Web sites serve narrow information topics and tend to form communities through hyperlinks (Chakrabarti et al., 1999).

Distributed IR has become a fast-growing research topic in recent years. Recent distributed IR research has been focused on intra-system retrieval fusion, cross-system communication, decentralized P2P network, and distributed information storage and retrieval algorithms (Callan, Crestani, & Sanderson, 2003). Research also concentrated on genetic algorithms for feature selection, intelligent crawling, information routing, etc. By now, modeling agent collaboration for text classification has drawn little attention from IR researchers. However, related work, although limited, has shown potential impact in this area. Peng, Mukhopadhyay, Raje, Palakal, and Mostafa (2001) compared single-agent (i.e. centralized) and multi-agent classifiers and pointed out that multi-agent classification has several advantages such as fault tolerance, adaptability, flexibility, resource sharing, privacy, and economics.

In this paper, we developed a multi-agent framework where agents had limited/distributed knowledge for text classification and collaborated with each other to overcome the knowledge distribution. Each agent was equipped with a certain learning algorithm for predicting potential collaborators and/or helping agents. We conducted experimental research on a standard news collection to examine the impact of two learning algorithms: the Pursuit Learning and the Nearest Centroid Learning. On a basic retrieval operation, namely classification, both algorithms achieved competitive classification effectiveness and efficiency. The impact of the learning exploration rate and the maximum collaboration range on classification effectiveness and efficiency were examined as well. Close investigation of agent learning dynamics revealed increasing and stabilizing patterns that were enhanced by the learning algorithms. We conclude the paper by describing future experimental research in this area.

References

- Börner, K., Dall'Asta, L., Ke, W., & Vespignani, A. (2005, April). Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity, special issue on Understanding Complex Systems*, 10(4), 58-67.
- Callan, J., Crestani, F., & Sanderson, M. (2003, September). Sigir 2003 workshop reports: Sigir 2003 workshop on distributed information retrieval. *ACM SIGIR Forum*, 37(2).
- Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Kumar, S., Raghavan, P., et al. (1999, August). Mining the link structure of the world wide web. *IEEE Computer*.
- Knight, K. (1999). Mining online text. *Commun. ACM*, 42(11), 58-61.
- Lewis, D. D. (1992). Text representation for intelligent text retrieval: a classification-oriented view. In *Text-based intelligent systems: current research and practice in information extraction and retrieval* (pp. 179-197). Hillsdale, NJ: Lawrence Erlbaum.
- Mostafa, J., & Lam, W. (2000). Automatic classification using supervised learning in a medical document filtering application. *Information Processing & Management*, 36(3), 415-444.
- Peng, S., Mukhopadhyay, S., Raje, R., Palakal, M., & Mostafa, J. (2001). A comparison between single-agent and multi-agent classification of documents. In *15th international parallel and distributed processing symposium* (p. 20090b).
- Red'ko, V. G. (2002). Problem of the global brain and multi-agent modeling. *icaais*, 00, 279.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1), 1-47.
- Taulbee, O. E. (1965). Invited papers-1: classification in information storage and retrieval. In *Proceedings of the 1965 20th national conference* (pp. 119-137). New York, NY, USA: ACM Press. (Chairman-R. W. House)
- Wimmer, M., & Traummuller, R. (2000). Trends in electronic government: Managing distributed knowledge. *dexa*, 00, 340.
- Yang, K. (2002). *Combining text-, link-, and classification-based retrieval methods to enhance information discovery on the web*. Unpublished doctoral dissertation, University of North Carolina at Chapel Hill.